Int. J. Advance Soft Compu. Appl, Vol. 10, No. 2, July 2018
ISSN 2074-8523

Prediction Models of Diabetes Diseases Based on Heterogeneous Multiple Classifiers

I Gede Agus Suwartane[1], Mohammad Syafrullah[1], Krisna Adiyarta[1]

[1]Program Studi Magister Ilmu Komputer, Universitas Budi Luhur, Jakarta, Indonesia
e-mail: agus.suwartane@gmail.com

Abstract

Diabetes disease is one of the global and most important health problems of the 21st century with the number of patients growing every year. One in two diabetics is undiagnosed patients, consequently many patients who already have severe complications. One way to reduce and slow the complications of diabetes is to make an early diagnosis. With the development of data mining, developed various models predicting diabetes by using data mining techniques. The main problem in building predictive models is how to improve the accuracy of predicted results. In this research, Heterogeneous Multiple Classifiers diabetic prediction model is developed by combining Support Vector Machine (SVM), K- Nearest Neighbor (KNN) and Decision Tree (C4.5) using Majority Voting. The prediction model based on Heterogeneous Multiple Classifiers was constructed to produce 93.56% accuracy, 97.48% sensitivity, 89.22% specificity, 91.16% precision and 94.13% F-Measure. The resulting performance value of Heterogeneous Multiple Classifiers based prediction model is higher than the performance value of Single Classifier-based prediction model used in building prediction model based on Heterogeneous Multiple Classifiers. Optimization conducted on prediction model based on Heterogeneous Multiple Classifiers in this study also proved to improve the performance of the prediction model.

Keywords: *Diabetes, Heterogeneous Multiple Classifiers, SVM, KNN, C4.5*

1 Introduction

Diabetes is a chronic disease that occurs because the pancreas cannot produce enough insulin or when the bodies can no longer use insulin effectively [1, 2]. This disease can lead to various complications including liver damage, heart, kidney, and blindness that can reduce productivity, disability and eventually can lead to premature death. One way to prevent or slow long-term complications for undiagnosed diabetes is to make an early diagnosis [2]. The main problem in diagnosis is the degree of accuracy of the diagnosis.

The use of data mining in the world of health is needed to produce a tool that is very useful, effective and fast in analyzing and obtaining important information from existing health data [3]. Data mining has tremendous potential for exporting hidden patterns in a large collection of medical record data. These patterns can later be used to diagnose the disease. This is what causes the use of data mining can help in early detection of a disease [4]. In data mining there are various kinds of training algorithms used to build prediction models. The application of a single classifier algorithm, such as SVM, C4.5, ANN, KNN and Naïve Bayes, in building predictive models to predict disease has been widely used and developed in studies. Prediction models built on single classifier produce good accuracy but these predictive models still have limitations to uncertainty [5].

Limitations of single classifier implementation to achieve the optimal level of accuracy led researchers to develop research toward the development of predictive models using multiple training algorithms (multiple classifiers). The application of multiple classifiers aims to make the predictive model generated to achieve a more accurate level of accuracy for various conditions [6]. This is possible because multiple classifiers can synthesize single classifier prediction results by using certain combination methods to improve the prediction accuracy [5].

In this study built a prediction model based on Heterogeneous Multiple Classifiers that can be used as an alternative to predict diabetes. The type of diabetes that is predicted in this study is type 2 diabetes. Prediction of diabetes is done in this study based on clinical symptoms. The research data used is patient medical record data at Lakespra Saryanto Jakarta. The prediction model is based on Heterogeneous Multiple Classifiers using 3 different training algorithms, namely Support Vector Machine (SVM), K - Nearest Neighbor (KNN) and Decision Tree (C4.5) with Majority Voting method. This study aims to find out how the performance of the predicted models of diabetes-based Heterogeneous Multiple Classifiers and how to optimize the prediction model of diabetes-based Heterogeneous Multiple Classifiers to improve the value of accuracy produced.

2 Related Work

2.1 Data Mining

Data mining is the process of extracting knowledge from large volumes of data stored in the database, data warehouse, or information stored in the repository. Data mining is also defined as a process of exploring new knowledge, a pattern that is divided from the large amount of data stored in the repository or storage by using pattern recognition techniques as well as statistics and mathematical techniques. Data mining is the core of the Knowledge Discovery in Database (KDD) process, which is an organized process for identifying valid, new, useful, and understandable patterns from a large and complex dataset. Steps in the Knowledge Discovery in Database (KDD) process include creating an application domain understanding, selecting and creating data sets where the process of knowledge discovery will be performed, preprocessing and cleansing, transforming data, selecting suitable data mining tasks, selecting data mining algorithms, data mining algorithms, evaluation and use of acquired knowledge [7].

2.2 Support Vector Machine (SVM)

Support Vector Machine (SVM) is rooted in the theory of statistical learning. SVM works well on high-dimensional data sets, even SVMs using kernel techniques should map original data from their original dimensions to other relatively higher dimensions. In SVM only a select number of data contribute to form the model used in the classification. SVM keeps only a small part of the training data to be used at the time of prediction. The data that contribute is called support vector. The basic idea of SVM is to maximize hyper plane boundaries. The concept of classification with SVM as an attempt to find the best hyper plane that serves as a separator of two classes of data on the input space. The best separator hyper plane between the two classes can be found by measuring the hyper plane's margins and searching for the maximum point. Margin is the distance between the hyper plane and the closest data from each class. The closest data is called support vector. The effort to locate this hyper plane is at the heart of the training process on SVM [8].

2.3 K - Nearest Neighbor (KNN)

The NN method is included in the lazy leaner classification, because it delayed the training process (or even did nothing at all) until there is a test data that the class label wants to know, and then the new method will run its algorithm. NN algorithm classifies based on similarity of data with other data. The closer the training data location to the test data, it can be said that the trainer data which is more considered similar to the test data. The smaller the value of the incapacity (distance) the more likely the test data with a number of neighbors K. Because the K wants the nearest neighbor then the metric used in the selection of the nearest

neighbor's K is the size of the incapacity (distance). The K value used in KNN states the number of nearest neighbors involved in predicting class label labels on the test data. From the nearest neighbor K is selected then a class voting from the nearest neighbor K is. The class with the largest number of neighboring votes is assigned as predicted class label on the test data [8].

2.4 C4.5

The famous decision tree algorithm is C4.5 which is the development of the ID3 algorithm. Algorithm C4.5 has advantages that are easy to understand, flexible and interesting because it can be visualized in the form of images. This algorithm is a tree structure where there is a node that describes the attributes, each branch describes the result of the attribute tested and each leaf represents the class. The C4.5 algorithm recursively visits each decision node, and chooses the optimal division until it cannot be subdivided. The C4.5 algorithm uses the concept of Information Gain or Entropy Reduction to select the optimal division [9].

2.5 Multiple Classifiers

Multiple Classifiers System (MCS) is one category of Hybrid Intelligent System (HIS). Multiple Classifiers System is a combination of several classifier both based on Homogeneous (kind) and Heterogeneous (different type) to give decision result [10]. Multiple classifiers system can also be defined as a set of single classifiers with their respective predictions combined with a way (fuser) to determine the classification of new objects.

Ensemble Design leads to how the characteristics of a classifier are complementary to achieve high levels of accuracy and diversity. Fuser Design can be developed with reference to Class Label Fusion (Unanimous Voting, Simple Majority and Majority Voting), Support Function Fusion and Trainable Fuser.
The two approaches used to build Multiple Classifiers are:

1. Homogeneous Multiple Classifiers, using the same classifier to process different data inputs.
2. Heterogeneous Multiple Classifiers, using different classifier-classifier to process the same data input.

In general the structure of the Multiple Classifiers System can be seen in Fig. 1.

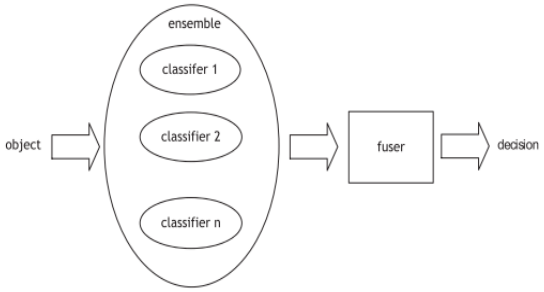

Fig. 1: Structure of Multiple Classifiers System

2.6 Performance Measurement Classifier

Classifier performance includes accuracy, sensitivity/recall, specificity, precision and F-Measure. Generally how to measure classification performance is using the Confusion Matrix as shown in Table 1.

Table 1: Confusion Matrix

		Predicted Results Class	
		Positive	**Negative**
Original Class	**Positive**	*True Positive* (TP)	*False Negative* (FN)
	Negative	*False Positive* (FP)	*True Negative* (TN)

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} \quad (1)$$

$$Sensitivity\ /Recall = \frac{TP}{TP+FN} \quad (2)$$

$$Specificity = \frac{TN}{FN+TN} \quad (3)$$

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

$$F - Measure = 2\ x\ \frac{Precision\ x\ Recall}{Precision+Recall} \quad (5)$$

2.7 Study Overview

Several studies have been done to build a prediction model of Heterogeneous Multiple Classifiers-based disease:

1. Najmeh Hosseinpour, et al. did research to diagnose diabetes by using several classifiers such as Bayesian, Functional, Rule-Base, Decision Trees and Bagging Ensemble. The dataset used is Pima Indian Diabetes. The results of this study indicate that Bagging Ensemble Classifiers have a better level of accuracy than other classifier. Accuracy value achieved by 77.47% [11].
2. Saba Bashir, Usman Qamar and Farhan Hassan Khan conducted research to diagnose Breast Cancer disease. The study used four BCD datasets, WDBC, Wisconsin, WPBC, taken from UCI and the Wisconsin Clinical Science Center. This study combined 5 classifier namely Naïve Bayes, Decision Tree Gini Index, Decision Tree Information Gain, Support Vector Machine and Memory Based Leaner and use Fisher Score as feature selection. The merging method uses Weighted Majority Voting. The results of this study obtained the average Accuracy = 85.23%, average Precision = 86.18% and Recall average = 76.68% [6].
3. Sadri Sa'di, et al. doing research by comparing the performance of several classifiers such as Naïve Bayes, RBF Network and J48 in diagnosing diabetes. The dataset used is Pima Indians Dataset. This study yields the conclusion that Naïve Bayes has the highest degree of accuracy among other classifiers. The resulting accuracy is 76.95% [12].
4. Harsha Sethi, et al. conducted research by combining ANN, Naïve Bayes, SVM and KNN by using Majority Voting to diagnose diabetes. The dataset used is 400 people with 10 attributes. The results of this study obtained an accuracy of 98.60% [13].

3 Problem Formulations or Methodology

3.1 Research Methods

The research method used is experimental method. The study begins with the collection of patient medical record data, preprocessing data, predictive model-building models using SVM, KNN and C4.5 and prediction models based on Heterogeneous Multiple Classifiers using SVM, KNN, C4.5 combination and combining the results with Simple Majority Vote as a result of the final prediction. The next step is to measure the performance of four prediction models built and compare their performance results. The next step is to optimize by choosing features and adjustment parameters of single classifier model, builder of prediction model based on Heterogeneous Multiple Classifiers, to get better accuracy.

3.2 Sampling

The data used in this study is the result of health examination data of patients who have checked their health and recorded in Lakespra Saryanto Jakarta from January

2015 until March 2016. Sampling method used in this study is Stratified Random Sampling. The number of samples used in these study as many as 450 samples consisted of 213 samples of patients who were diagnosed with diabetes and 237 samples of patients who were diagnosed not having diabetes.
The data in this study consist of 9 features and 2 classes, namely Age, Fasting Blood Sugar, Blood Sugar after 2 Hours Fasting, Cholesterol, Body Mass Index, Systolic Blood Pressure, Diastolic Blood Pressure, Abdominal Radiance and Diabetes.

3.3 Instrumentation

Hardware instruments used in the form of computers with specifications Intel Core 2 Duo processor, 2 GB RAM, P7570 2.26 Ghz. The software instrument used in this research is WEKA (Waikato Environment for Knowledge Analysis) version 3.6.4 to perform feature selection and MATLAB (Matrix Laboratory) R2015a version to build prediction model based on Single Classifier (SVM, KNN and C4.5) and predictive models based on Heterogeneous Multiple Classifiers.

4 The Proposed Method

The predictive model built consists of four models, three predictive models based on Single Classifier (prediction model using SVM, prediction model using KNN, prediction model using C4.5) and prediction model based on Heterogeneous Multiple Classifiers. The design prediction model of diabetes-based Heterogeneous Multiple Classifiers to be built in this study can be seen in Fig. 2. Each prediction model will be validated by using 10 fold cross validation to measure the performance.

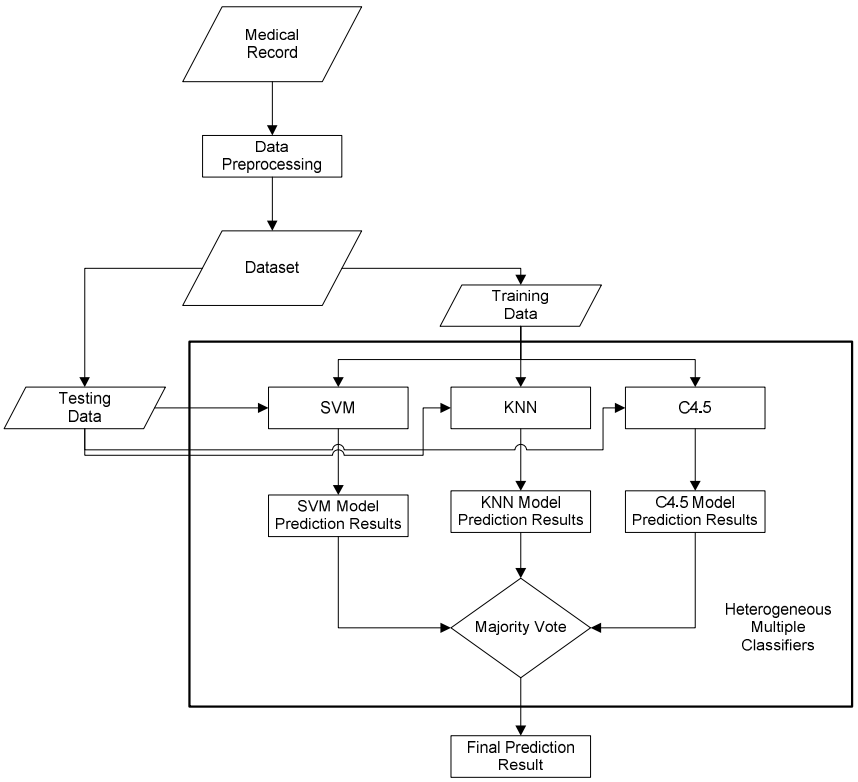

Fig. 2: Diabetic disease prediction model based on Heterogeneous Multiple Classifiers

5 Results, Analysis and Discussions

5.1 Research Data

Descriptive statistics of data used in this study can be seen in Table 2.

Table 2: Descriptive research data

Features	Minimum	Maximum	Average	Description
UM	22.00	85.00	47.17	Age
GP	72.00	434.00	142.61	Fasting Blood Sugar
G2	62.00	599.00	171.78	Blood Sugar After 2 Hours Fasting
KL	112.00	434.00	214.42	Cholesterol

BM	15.73	40.65	26.38	Body Mass Index
TS	90.00	180.00	120.53	Systolic Blood Pressure
TD	70.00	110.00	80.60	Diastolic Blood Pressure
LP	63.00	123.00	87.95	Abdominal Radiance
DM	Yes = 213 ; No = 237			Diagnosis Results

The data used in this study will be divided into 2 parts, namely training data and target data. Train data will be used to construct a classification model as a prediction model. Test data will be used to test the classification model that is formed. This data sharing is based on 10 fold cross validation as validation in this research, so the data will be divided into 10 groups with each group amounted to 45 pieces of data. Each iteration, 9 groups of data (405 data) will be used as training data and the rest will be used as test data.

5.2 Development and Performance Measurement Model Prediction

5.2.1 Prediction Model with C4.5 Method

The result of measurement accuracy, sensitivity / recall, specificity, precision and F-Measure prediction model with C4.5 method can be seen in Table 3.

Table 3: Performance model with C4.5 method

Iteration	Accuracy	Sensitivity	Specificity	Precision	F-Measure
1	88.89%	90.48%	87.50%	86.36%	88.37%
2	88.89%	95.24%	83.33%	83.33%	88.89%
3	86.67%	90.48%	83.33%	82.61%	86.36%
4	91.11%	100.00%	83.33%	84.00%	91.30%
5	93.33%	85.71%	100.00%	100.00%	92.31%
6	84.44%	85.71%	83.33%	81.82%	83.72%
7	100.00%	100.00%	100.00%	100.00%	100.00%
8	91.11%	95.45%	86.96%	87.50%	91.30%
9	91.11%	95.45%	86.96%	87.50%	91.30%
10	93.33%	95.45%	91.30%	91.30%	93.33%
Average	90.89%	93.40%	88.61%	88.44%	90.69%

5.2.2 Prediction Model with KNN Method

The result of measurement accuracy, sensitivity / recall, specificity, precision and F-Measure prediction model with KNN method can be seen in Table 4.

Table 4: Performance model with KNN method

Iteration	Accuracy	Sensitivity	Specificity	Precision	F-Measure
1	95.56%	100.00%	90.48%	92.31%	96.00%
2	95.56%	100.00%	90.48%	92.31%	96.00%
3	86.67%	91.67%	80.95%	84.62%	88.00%
4	97.78%	95.83%	100.00%	100.00%	97.87%
5	93.33%	100.00%	85.71%	88.89%	94.12%
6	86.67%	95.83%	76.19%	82.14%	88.46%
7	95.56%	100.00%	90.48%	92.31%	96.00%
8	93.33%	100.00%	86.36%	88.46%	93.88%
9	97.78%	100.00%	95.45%	95.83%	97.87%
10	84.44%	95.65%	72.73%	78.57%	86.27%
Average	92.67%	97.90%	86.88%	89.54%	93.45%

5.2.3 Prediction Model with SVM Method

The result of measurement accuracy, sensitivity / recall, specificity, precision and F-Measure prediction model with SVM method can be seen in Table 5.

Table 5: Performance model with SVM method

Iteration	Accuracy	Sensitivity	Specificity	Precision	F-Measure
1	95.56%	100.00%	90.48%	92.31%	96.00%
2	95.56%	100.00%	90.48%	92.31%	96.00%
3	93.33%	95.83%	90.48%	92.00%	93.88%
4	95.56%	91.67%	100.00%	100.00%	95.65%
5	91.11%	95.83%	85.71%	88.46%	92.00%
6	88.89%	95.83%	80.95%	85.19%	90.20%
7	95.56%	100.00%	90.48%	92.31%	96.00%
8	93.33%	100.00%	86.36%	88.46%	93.88%
9	95.56%	95.65%	95.45%	95.65%	95.65%
10	88.89%	95.65%	81.82%	84.62%	89.80%
Average	93.33%	97.05%	89.22%	91.13%	93.91%

5.2.4 Prediction Model with Heterogeneous Multiple Classifiers (HMC) Method

The result of measurement accuracy, sensitivity / recall, specificity, precision and F-Measure prediction model with HMC method can be seen in Table 6.

Table 6: Performance model with HMC method

Iteration	Accuracy	Sensitivity	Specificity	Precision	F-Measure
1	95.56%	100.00%	90.48%	92.31%	96.00%
2	95.56%	100.00%	90.48%	92.31%	96.00%

3	88.89%	91.67%	85.71%	88.00%	89.80%
4	95.56%	91.67%	100.00%	100.00%	95.65%
5	93.33%	100.00%	85.71%	88.89%	94.12%
6	88.89%	95.83%	80.95%	85.19%	90.20%
7	97.78%	100.00%	95.24%	96.00%	97.96%
8	93.33%	100.00%	86.36%	88.46%	93.88%
9	97.78%	100.00%	95.45%	95.83%	97.87%
10	88.89%	95.65%	81.82%	84.62%	89.80%
Average	93.56%	97.48%	89.22%	91.16%	94.13%

5.2.5 Comparison of Predictive Model Performance

Comparison of predictive model performance of each method can be seen in Table 7.

Table 7: Comparison of predictive model performance

Performance Measures	Method			
	C4.5	KNN	SVM	HMC
Accuracy	90.89%	92.67%	93.33%	93.56%
Sensitivity / Recall	93.40%	97.90%	97.05%	97.48%
Specificity	88.61%	86.88%	89.22%	89.22%
Precision	88.44%	89.54%	91.13%	91.16%
F-Measure	90.69%	93.45%	93.91%	94.13%

Based on the performance measurement results of each prediction model can be seen that the accuracy, specificity, precision and F-Measure value generated by Heterogeneous Multiple Classifiers based prediction model is higher than all Single Classifier-based models. The value of sensitivity / recall generated by Heterogeneous Multiple Classifiers based prediction model is still lower than the prediction model by KNN method, but higher than the prediction model with C4.5 and SVM.

5.3 Optimization of HMC-Based Prediction Model

5.3.1 Feature Selection

Feature selection is done because in this study, allegedly able to optimize prediction model based on Heterogeneous Multiple Classifiers. Feature selection is done by using InfoGainAttributeEvaluator as feature evaluation and Ranker as Search Method. The result of feature selection can be seen in Fig. 3.

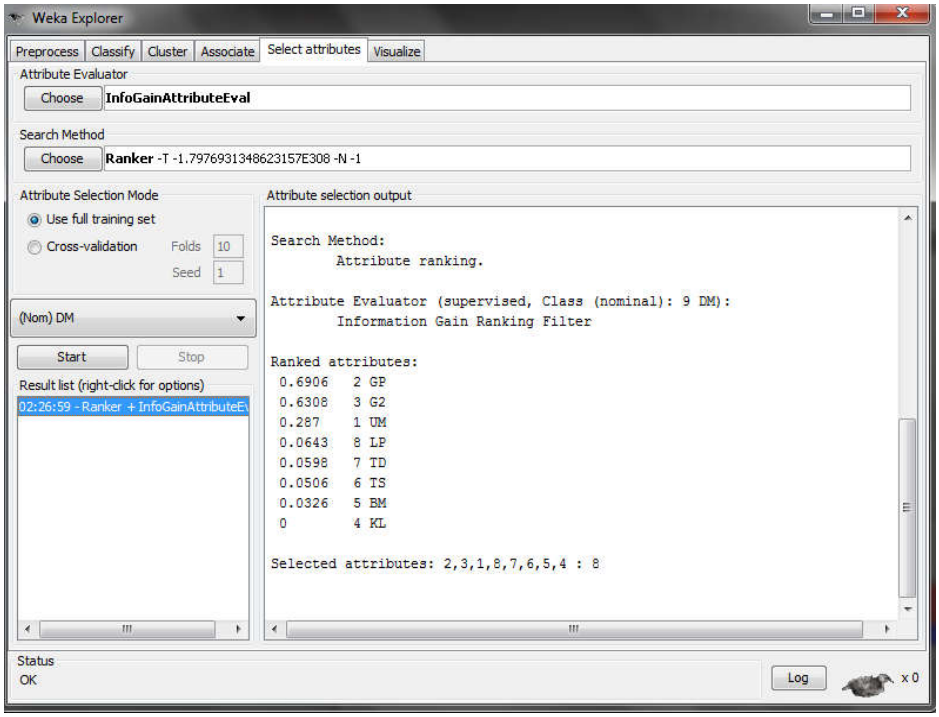

Fig 3: Selection of feature results with WEKA

The sequences of features with the highest rankings are: GP, G2, UM, LP, TD, TS, BM and KL. Next is to build predictive model by removing one by one feature based on the results of feature selection.

5.3.2 Processing with Features [UM, GP, G2, BM, TS, TD, LP]

Establishment of prediction models using UM, GP, G2, BM, TS, TD and LP features where the KL features are omitted, giving results for each model that can be seen in Table 8.

Table 8: Performance results with features reduction KL

Performance Measures	Method			
	C4.5	KNN	SVM	HMC
Accuracy	90.67%	92.00%	93.11%	93.11%
Sensitivity / Recall	92.49%	96.63%	97.05%	96.63%
Specificity	89.02%	86.86%	88.72%	89.20%
Precision	88.85%	89.43%	90.84%	91.10%
F-Measure	90.43%	92.79%	93.74%	93.70%

5.3.3 Processing with Features [UM, GP, G2, TS, TD, LP]

Establishment of predictive models using UM, GP, G2, TS, TD and LP features, where KL and BM are omitted, give results for each model as seen in Table 9.

Table 9: Performance results with features reduction KL, BM

Performance Measures	Method			
	C4.5	KNN	SVM	HMC
Accuracy	90.89%	91.56%	93.11%	92.89%
Sensitivity / Recall	91.54%	96.21%	97.05%	96.21%
Specificity	90.29%	86.39%	88.72%	89.20%
Precision	89.76%	89.09%	90.84%	91.10%
F-Measure	90.50%	92.37%	93.74%	93.47%

5.3.4 Processing with Features [UM, GP, G2, TD, LP]

Establishment of predictive models using UM, GP, G2, TD and LP features where KL, BM and TS are omitted, giving results for each model as seen in Table 10.

Table 10: Performance results with features reduction KL, BM, TS

Performance Measures	Method			
	C4.5	KNN	SVM	HMC
Accuracy	90.89%	92.44%	92.67%	93.11%
Sensitivity / Recall	91.54%	97.05%	97.46%	97.05%
Specificity	90.29%	87.32%	87.32%	88.74%
Precision	90.05%	89.83%	89.94%	90.82%
F-Measure	90.58%	93.16%	93.42%	93.73%

5.3.5 Processing with Features [UM, GP, G2, LP]

Establishment of predictive models using UM, GP, G2 and LP features in which KL, BM, TS and TD features are omitted, giving results for each model as seen in Table 11.

Table 11: Performance results with features reduction KL, BM, TS, TD

Performance Measures	Method			
	C4.5	KNN	SVM	HMC
Accuracy	91.78%	92.00%	93.56%	92.89%
Sensitivity / Recall	92.92%	96.20%	97.48%	96.63%
Specificity	90.71%	87.34%	89.20%	88.72%
Precision	90.81%	89.84%	91.22%	90.78%
F-Measure	91.59%	92.71%	94.16%	93.49%

5.3.6 Processing with Features [UM, GP, G2]

Establishment of prediction models using UM, GP, G2 features in which features KL, BM, TS, TD and LP are omitted, giving results for each model seen in Table 12.

Table 12: Performance results with features reduction KL, BM, TS, TD, and LP

Performance Measures	Method			
	C4.5	KNN	SVM	HMC
Accuracy	92.89%	92.67%	93.11%	92.67%
Sensitivity / Recall	92.47%	97.45%	96.65%	95.78%
Specificity	93.26%	87.34%	89.18%	89.18%
Precision	92.62%	89.86%	91.20%	91.10%
F-Measure	92.44%	93.41%	93.72%	93.26%

5.3.7 Processing with Features [GP, G2]

Establishment of prediction models using GP and G2 features in which the features of KL, BM, TS, TD, LP and MW are omitted, giving results for each model seen in Table 13.

Table 13: Performance results with features reduction KL, BM, TS, TD, LP, UM

Performance Measures	Method			
	C4.5	KNN	SVM	HMC
Accuracy	90.89%	92.22%	92.22%	92.22%
Sensitivity / Recall	91.06%	97.05%	96.63%	96.63%
Specificity	90.69%	86.86%	87.32%	87.32%
Precision	90.15%	89.50%	89.89%	89.85%
F-Measure	90.44%	93.01%	92.98%	93.00%

5.3.8 Feature Selection Analysis

From the tables described above, it can be seen that the accuracy value for the prediction model using C4.5 method reaches the highest value in data processing with UM, GP and G2 features of 92.89%. This value is higher than the value of accuracy when processing the data with all the features that is 90.89%.

The highest accuracy value for the prediction model by KNN method is achieved on data processing with features of UM, GP and G2 of 92.67%. This value is equal to the value of accuracy when processing data with all features.

The highest accuracy score for predictive model with SVM method was achieved on data processing with UM, GP, G2 and LP features of 93.56%. This value is higher than the accuracy value when processing the data with all features of

93.33%.

For Heterogeneous Multiple Classifiers-based prediction models, the highest accuracy score of 93.11% was obtained by reducing the features of KL, BM and TS. Accuracy value obtained is still smaller than the value of accuracy generated when processing all the features of 93.56%.

5.3.9 Parameter Change

To obtain a better accuracy value, parameter changes were used to construct a prediction model based on Heterogeneous Multiple Classifiers. The parameter changes made can be seen in Table 14.

Table 14: Parameter change method of classification

Method	Parameter	Before	After
SVM	Kernel Function	'Linear'	'Quadratic'
	Method	'SMO'	'LS'
KNN	K (K Nearest Neighbor)	10	30

Values of accuracy, sensitivity / recall, specificity, precision, F-Measure generated prediction model based on Heterogeneous Multiple Classifiers after change of parameter of classification method, can be seen in Table 15.

Table 15: HMC model performance after optimization

Iteration	Accuracy	Sensitivity	Specificity	Precision	F-Measure
1	95.56%	100.00%	90.48%	92.31%	96.00%
2	95.56%	100.00%	90.48%	92.31%	96.00%
3	91.11%	95.83%	85.71%	88.46%	92.00%
4	95.56%	91.67%	100.00%	100.00%	95.65%
5	93.33%	100.00%	85.71%	88.89%	94.12%
6	86.67%	95.83%	76.19%	82.14%	88.46%
7	97.78%	100.00%	95.24%	96.00%	97.96%
8	95.56%	100.00%	90.91%	92.00%	95.83%
9	97.78%	100.00%	95.45%	95.83%	97.87%
10	88.89%	95.65%	81.82%	84.62%	89.80%
Average	93.78%	97.90%	89.20%	91.26%	94.37%

While the comparison of performance prediction model based on Heterogeneous Multiple Classifiers before optimization and after optimization can be seen in Table 16.

Table 16: HMC model performance comparison before and after optimization

Performance Measures	Before Optimization	After Optimization
Accuracy	93.56%	93.78%
Sensitivity / Recall	97.48%	97.90%
Specificity	89.22%	89.20%
Precision	91.16%	91.26%
F-Measure	94.13%	94.37%

From the table can be seen that the value of accuracy increased by 0.22%. The sensitivity / recall rate increased by 0.42%, the precision value increased by 0.10%, and the F-Measure value increased by 0.24% but the specificity value decreased by 0.02%.

6 Conclusion

The prediction model based on Heterogeneous Multiple Classifiers built in this study resulted in accuracy of 93.56%, sensitivity / recall of 97.48%, specificity 89.22%, precision 91.16% and F-Measure of 94.13 %. This result is higher than the results achieved by Single Classifiers-based prediction models (C4.5, KNN and SVM) used in building predictive models based on Heterogeneous Multiple Classifiers. This study also shows that feature selection and parameter changes on SVM and KNN methods can optimize prediction model based on Heterogeneous Multiple Classifiers by increasing accuracy value to 93,78%, sensitivity / recall become 97,90%, precision become 91,26% and F -Measure to 94.37%. For further research the Heterogeneous Multiple Classifiers based prediction model generated from this study can be tested on another dataset. The performance of this prediction model can also be compared with the performance of predictive models built using other classification methods and fuser methods.

References

[1] World Health Organization, *Global Report on Diabetes*, vol. 978. 2016.

[2] IDF, *IDF Diabetes Atlas*. 2015.

[3] D. Tomar and S. Agarwal, “A survey on Data Mining approaches for Healthcare,” vol. 5, no. 5, pp. 241–266, 2013.

[4] A. S. Gowri, “Hybrid Intelligent System of Heterogeneous Classifiers for Breast Cancer Diagnosis” vol. 6, no. 1, pp. 30–35, 2016.

[5] J. Sun and H. Li, “Listed companies’ financial distress prediction based on weighted majority voting combination of multiple classifiers,” vol. 35, pp. 818–827, 2008.

[6] S. Bashir, U. Qamar, and F. Hassan, "Heterogeneous classifiers fusion for dynamic breast cancer diagnosis using weighted vote based ensemble" *Qual. Quant.*, pp. 2061–2076, 2015.

[7] O. Maimon and L. Rokach, *Data Mining and Knowledge Discovery Handbook*, 2nd ed. New York: Springer Science + Bussiness Media, 2010.

[8] E. Prasetyo, *Data Mining Mengolah Data Menjadi Informasi Menggunakan Matlab*, I. Yogyakarta: Penerbit Andi, 2014.

[9] P. P. Widodo, R. T. Handayanto, and Heriawati, *Penerapan Data Mining dengan Matlab*, 1st ed. Bandung: Penerbit Rekayasa Sains, 2013.

[10] M. Wozniak, M. Graña, and E. Corchado, "A survey of multiple classifier systems as hybrid systems" vol. 16, pp. 3–17, 2014.

[11] N. Hosseinpour and K. Ansari-asl, "Diabetes Diagnosis by Using Computational Intelligence Algorithms" vol. 2, no. 12, pp. 71–77, 2012.

[12] S. Sa'di, A. Maleki, R. Hashemi, Z. Panbechi, and K. Chalabi, "Comparison of Data Mining Algorithms in the Diagnosis of Type II Diabetes" *Int. J. Comput. Sci. Appl.*, vol. 5, no. 5, pp. 1–12, 2015.

[13] H. Sethi, "Artificial Intelligence based Ensemble Model for Diagnosis of Diabetes" vol. 8, no. 4, pp. 1540–1548, 2017.